Responsible AI Governance Maturity Model
Full Questionnaire & Scoring Guidelines

Based on the NIST AI RMF   Full paper: Dotan et al. (2024)
Supported by the Notre Dame-IBM Tech Ethics Lab

At a Glance - Across 3 Development Phases

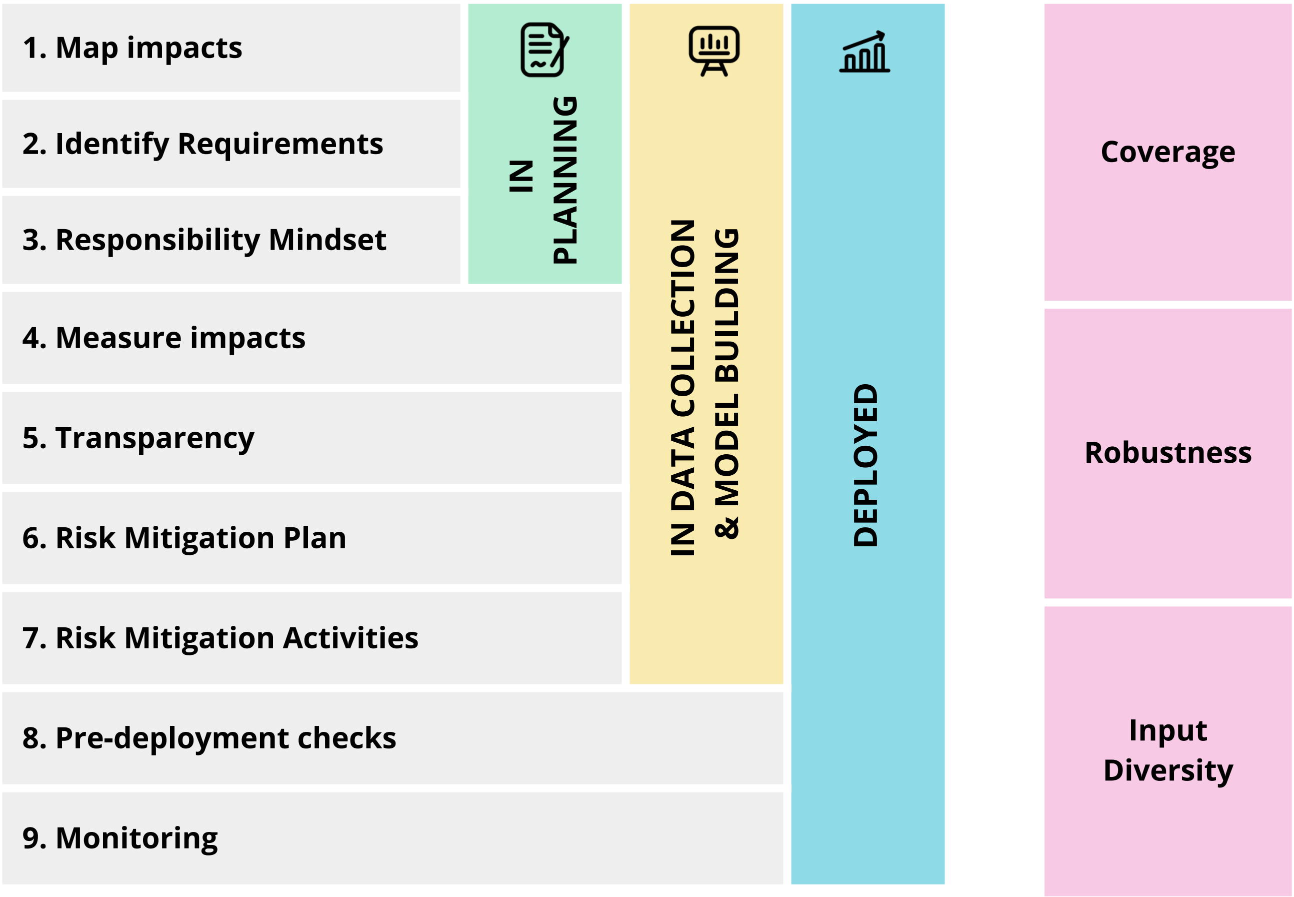
PLANNING

DATA COLLECTION & MODEL BUILDING

DEPLOYMENT

Questions - 9 Topics Across the 3 Phases

Metrics



1. Map impacts
2. Identify Requirements
3. Responsibility Mindset
4. Measure impacts
5. Transparency
6. Risk Mitigation Plan
7. Risk Mitigation Activities
8. Pre-deployment checks
9. Monitoring

IN PLANNING

IN DATA COLLECTION & MODEL BUILDING

DEPLOYED

Coverage

Robustness

Input Diversity

Scoring Guidelines: Metrics

The score of each topic should be based on the three metrics below. The evaluator ranks how well each metric is satisfied: Low, medium, or high.

Coverage	The scoring of each topic should be higher the better the coverage of the activities in the substatements.
Robustness	Scores should be higher the more the activities are robust. Activities that are robust share the following characteristics: • **Regularity** - Performed in a routine manner • **Systematicity** - Follow policies that are well-defined and span company-wide • **Trained Personnel** - Performed by people who are properly trained and whose roles in the activities are clearly defined • **Sufficient Resources** - Supported by sufficient resources, including budget, time, compute power, and cutting-edge tools • **Adaptivity** - Adapting to changes in the landscape and product, including regular reviews and effective contingency processes to respond to failure • **Cross-functionality** - All core business units and senior management are informed of the outcomes and contribute to decision-making, strategy, and resource allocation related to the activities (core business units include finance, customer support, HR, marketing, sales, etc.)
Input Diversity	Input diversity means that the activities are informed by input from diverse internal and external stakeholders: • **A low level of input diversity** means that the relevant activities receive input from relatively few kinds of stakeholders, such as members of one internal team only. • **High levels of input diversity** mean that the activities receive input from diverse internal and external stakeholders. For example, suppose that a company chooses its fairness metrics in consultation with civil society organizations, surveys of diverse customers administered by the customer success team, and conversations with diverse employees in the company. In that case, the company demonstrates a high level of input diversity with regard to the statement “We evaluate and document bias and fairness issues related to this AI system”.

Scoring Guidelines: Explanations

Scores must be accompanied with an explanation. The explanation should refer to information about what the organization does or doesn't do and any relevant contextual facts.

Resources - Where to find relevant information? (Examples)

Internal information (if available)	• Internal documents • Interviews with employees • Informal conversations and employee knowledge • Internal metrics, e.g., Objectives and Key Results (OKRs) and Key Performance Indicators (KPIs)
External information	• External company documents, e.g. AI ethics frameworks • External company reports, e.g., annual or ESG reports • Research papers by or about the company • Media reports • Lawsuits

Evaluation Elements - What to evaluate when scoring? (Examples)

Execution	Outcomes, procedures, and resources dedicated to activities: ○ RAI metrics and progress on those metrics ○ How much time is spent working on certain tasks ○ The execution of RAI best practices, such as red-team exercises or ethics reviews
Uptake	How relevant activities and deliverables are received: ○ Whether the outputs of RAI work are officially adopted by the company ○ Whether and how the company's leadership supports it
People	Who performs the relevant activities and how: ○ Whether the people conducting the relevant activities are doing so as part of their official capacity or as a voluntary side project ○ The number of people assigned to relevant tasks ○ How suitable these people are to perform the tasks
Communication	The nature of internal conversations related to the relevant activities: ○ How frequently relevant topics or tasks are discussed ○ The depth of the conversations ○ The formality of conversation channels for relevant topics. For example, is there a dedicated time to talk about it, or does it come up occasionally in Slack?

Scoring Guidelines: Overall Calculation

The overall score of each topic is calculated based on the scores of all the metrics. It ranges from 1-5, where 1 is the lowest and 5 is the highest.

Score	Conditions	Shorthand
5	All three metrics are satisfied to a high degree	HHH
4	Two of the metrics are satisfied to a high degree and one to a medium degree	HHM
3	One of the following is the case: • Two of the metrics are satisfied to a medium degree and one to a high degree • Two of the metrics are satisfied to a high degree and one to a low degree • One metric is satisfied to a high degree, one to a medium degree, and one to a low degree. • All three metrics are satisfied to a medium degree.	HMM HHL HML MMM
2	One of the following is the case: • Two of the metrics are satisfied to a medium degree and one to a low degree • One metric is satisfied to a medium degree and two to a low degree • One of the metrics is satisfied to a high degree and two to a low degree.	MML MLL HLL
1	All metrics are satisfied to a low degree	LLL

Full Questionnaire: All Phases

1. Planning: Map Impacts

We clearly define what the AI is supposed to do and its impacts, including scope, goals, methods, and negative and positive potential impacts of these activities.

1.1	**Goals**	We define the goals, scope, and methods of this AI system.
1.2	**Positive Impacts**	We identify the benefits and potential positive impacts of this AI system, including the likelihood and magnitude.
1.3	**Business Value**	We identify the business value of this AI system.
1.4	**Negative Impacts**	We identify the possible negative impacts of this AI system, including the likelihood and magnitude.
1.5	**Costs of Malfunction**	We identify the potential costs of malfunctions of this AI system, including non-monetary costs such as decreased trustworthiness.
1.6	**Unexpected Impacts**	We implement processes to integrate input about unexpected impacts.
1.7	**Methods and Tools**	We identify the methods and tools we use for mapping impacts.

Full Questionnaire: All Phases

2. Planning: Identify Requirements

We identify the requirements the AI must meet, including compliance, certifications, and human oversight needs.

2.1	**Human Oversight**	We identify the human oversight processes the system needs.
2.2	**Certifications**	We identify the technical standards and certifications the system will need to satisfy.
2.3	**Legal Requirements**	We identify AI legal requirements that apply to this AI system.

3. Planning: Responsibility Mindset

We facilitate a mindset of responsibility, for example, by providing AI ethics training to relevant personnel, clearly defining relevant roles, establishing policies, and implementing practices for critical thinking.

3.1	**Policies and Guidelines**	We write policies and guidelines about AI ethics.
3.2	**Roles and Responsibilities**	We document roles, responsibilities, and lines of communication related to AI risk management
3.3	**Training**	We provide training about AI ethics to relevant personnel.
3.4	**Critical Thinking**	We implement practices to foster critical thinking about AI risks.

Full Questionnaire: Data Collection & Model Building + Deployment Phases

4. Data Collection & Model Building: Measure Impacts

We measure potential negative impacts.

4.1	**Strategy for Measuring the Impacts**	We make and periodically re-evaluate our strategy for measuring the impacts of this AI system. It includes choosing which impacts we measure. It also includes how we will approach monitoring unexpected impacts and impacts that can't be captured with existing metrics.
4.2	**Methods and Tools**	We have a clear set of methods and tools to use when measuring the impacts of this AI system. It includes which metrics and datasets we use.
4.3	**Effectiveness**	We evaluate the effectiveness of our measurement processes.
4.4	**Performance**	We regularly revaluate and document the performance of this AI system in conditions similar to deployment.
4.5	**Bias and Fairness**	We regularly evaluate bias and fairness issues related to this AI system.
4.6	**Privacy**	We regularly evaluate privacy issues related to this AI system.
4.7	**Environmental**	We regularly evaluate environmental impacts related to this AI system.
4.8	**Transparency and Accountability**	We regularly evaluate transparency and accountability issues related to this AI system.
4.9	**Security and Resilience**	We regularly evaluate security and resilience issues related to this AI system
4.10	**Explainability**	We regularly evaluate explainability issues related to this AI system
4.11	**Third-party**	We regularly evaluate third-party issues, such as IP infringement, related to this AI system.
4.12	**Other Impacts**	We regularly evaluate other impacts related to this AI system.
4.13	**Human Subjects**	If evaluations use human subjects, they are representative and meet appropriate requirements.

Full Questionnaire: Data Collection & Model Building + Deployment Phases

5. Data Collection & Model Building: Transparency

We document information about the system, including explaining how it works, limitations, and risk controls.

5.1	**Human Oversight**	We document information about the system's limitations and options for human oversight related to this AI system. The documentation is good enough to assist those who need to make decisions based on the system's outputs.
5.2	**Risk Controls**	We document the system risk controls, including in third-party components.
5.3	**Model Explanation**	We explain the model to ensure responsible use.
5.4	**Inventory**	We inventory information about this AI system in a repository of our AI system.

6. Data Collection & Model Building: Risk Mitigation Plan

We plan how to respond to risks, including setting priorities and documenting residual risks.

6.1	**Plan**	We plan how we will respond to the risks caused by this AI system. The response options can include mitigating, transferring, avoiding, or accepting risks.
6.2	**Prioritization**	We prioritize the responses to the risks of this AI system based on impact, likelihood, available resources or methods, and the organization's risk tolerance.
6.3	**Residual Risks**	We identify the residual risks of this AI system (the risks that we do not mitigate). The documentation includes risks to buyers and users of the system.
6.4	**Unexpected Risks**	We have a plan for addressing unexpected risks related to this AI system as they come up.

Full Questionnaire: Data Collection & Model Building + Deployment Phases

7. Data Collection & Model Building: Risk Mitigation Activities

We act to minimize risks, including addressing your prioritized risks and tracking incidents.

7.1	**Meets Objectives**	We proactively evaluate whether this system meets its stated objectives and whether its development or deployment should proceed.
7.2	**Bias and Fairness**	We ensure this AI's bias and fairness performance meets our standards.
7.3	**Privacy**	We ensure this AI's privacy performance meets our standards.
7.4	**Environmental**	We ensure this AI's environmental performance meets our standards.
7.5	**Transparency and Accountability**	We ensure this AI's transparency and accountability meets our standards.
7.6	**Security and Resilience**	We ensure this AI's security and resilience meets our standards,
7.7	**Explainability**	We ensure this AI's explainability performance meets our standards.
7.8	**Third-party**	We ensure this AI's third-party impacts, such as IP infringement, meet our standards.
7.9	**Human Oversight**	We implement processes for human oversight related to this AI system.
7.10	**Appeal**	We implement processes for appeal related to this AI system.
7.11	**End-of-life Mechanisms**	We maintain end-of-life mechanisms to supersede, disengage, or deactivate this AI system if its performance or outcomes are inconsistent with the intended use.
7.12	**All Other Risks**	We address all other risks prioritized in our plans related to this system by conducting measurable activities.
7.13	**Unexpected Risks**	We address unexpected risks related to this system by conducting measurable activities.
7.14	**Errors and Incidents**	We track and respond to errors and incidents related to this system by conducting measurable activities.

Full Questionnaire: Deployment Phase

8. Deployment: Pre-Deployment Checks

We only release versions that meet our AI ethics standards.

8.1	**Valid and Reliable**	We demonstrate that this system is valid, reliable, and meets our standards. We document the conditions under which it falls short.

9. Deployment: Monitoring

We monitor and resolve issues as they arise.

9.1	**Monitoring Plan**	We plan how to monitor risks related to this system post-deployment.
9.2	**Functionality and Behavior**	We monitor this system's functionality and behavior post-deployment.
9.3	**Sustain Value**	We apply mechanisms to sustain the value of this AI system post-deployment.
9.4	**Input from Users**	We capture and evaluate input from users about this system post-deployment.
9.5	**Appeal and Override**	We monitor appeal and override processes related to this system post-deployment.
9.6	**Incidents and Response**	We monitor incidents related to this system and responses to them post-deployment.
9.7	**High-risk Third-party**	We monitor incidents related to high-risk third-party components and respond to them.
9.8	**All Other Components**	We implement all other components of our post-deployment monitoring plan for this system.
9.8	**End-of-life Mechanisms**	We monitor issues that would trigger our end-of-life mechanisms for this system, and we take the system offline if issues come up.